به نام خدا

علیرضا آخوندی
۹۷۳۱۱۰۷

سوال اول :
در تسک دسته بندی وظیفه مدل پیش بینی کردن برچسب برای یک داده ورودی است. در واقع خروجی این مدل‌ها گسسته است. در حالی که در رگرسیون خروجی پیوسته بوده و با توجه به ورودی مدل خروجی یک کمیت پیوسته و مقدار عددی است.

مثال های دسته بندی:
تشخیص بیماری: در این کاربرد، مدل تصاویر با فرمت های مختلف (MRI ، ...) دریافت کرده و مبتلا بوده یا عدم ابتلا به یک بیماری خاص را تشخیص می دهد.
تشخیص مانع در ماشین های خودران: در این کاربرد مدل یا با کمک داده های سنسور فاصله یاب یا با کمک تصویر گرفته شده از دوربین روبروی خودرو وجود یا عدم وجود یک مانع را تشخیص میدهد.
به دست آوردن خصوصیات چهره : برای این که بتوانیم خصوصیات چهره یک فرد را (مانند رنگ پوست، لبخند، عصبانی یا خوشحال بودن و ...) را به دست آوریم برای هر ویژگی میتوانیم یک دسته بند ایجاد کنیم.
Authentication: برای این که با کمک تصویر صورت یک فرد تشخیص دهیم که آیا آن فرد اجازه دسترسی به منابعی را (برای مثال ورود به آزمایشگاه) دارد می توانیم از دسته بند ها استفاده کنیم.
وفاداری: اینکه متوجه شویم مشتری یک شرکت خاص به آن شرکت وفادار است یا خیر بر اساس ویژگی های مختلف آن مشتری مانند: میزان موجودی، میزان خرید های آن مشتری، میزان استفاده آن مشتری از شرکت و...

مثال های رگرسیون:
پیش بینی سرعت خودرو: مدل با دریافت چند فریم از دوربین جلوی خودرو سرعت آن را پیش بینی میکند.
کاربرد های پزشکی: با توجه به سوابق بیمار می توان مقدار قند خون یا هر ویژگی دیگری را پیشبینی کرد.
معدل یک دانشجو: با توجه به سوابق دانشجو و روحیات او می توان معدل ترم او را پیش بینی کرد.
میزان مصرف یک دارو: برای درمان یک بیماری، میتوان میزان مصرف یک دارو را برای هر بیمار با توجه به سوابق او تنظیم کرد.
سهام : پیش بینی نوسانات سهام با کمک داده های از قبل جمع شده.

سوال دوم: بخش اول:

Accuracy: دقت یک مدل برابر است با تعداد داده هایی که درست طبقه بندی شده اند تقسیم بر تعداد کل داده ها و فرمول آن نیز به صورت زیر است:

$$Accuracy = \frac{TN + TP}{TN + FP + TP + FN}$$

Recall: نشان می دهد چند درصد از داده هایی که واقعا مربوط به یک دسته هستند در آن دسته قرار گرفته اند که فرمول آن به صورت زیر است:

$$Recall = \frac{TP}{TP + FN}$$

Precision: همان دقت است که نسبت پاسخ های مثبت درست را به پاسخ های مثبت مشخص میکند که به صورت زیر محاسبه میشود:

$$Precision = \frac{TP}{TP + FP}$$

F1-Score: حاصل ضرب precision و recall تقسیم بر حاصل جمع آن ها ضرب در دو.

بخش دوم :

- سناریو اول: فرض کنید که می خواهیم عملکرد مدلی را ارزیابی کنیم که وظیفه آن تشخیص یک نوع سرطان است. فرض کنید که در مجموعه داده ما داده های مربوط به سوابق 1000 بیمار قرار دارد و تنها 10 مورد آنها این سرطان را دارند. در این صورت اگر از معیار accuracy استفاده کنیم آنگاه مدلی که به ازای تمام سوابق پاسخ منفی می دهد، دقت 99 درصد خواهد داشت که دقت خوبی است. اما در عمل این مدل قادر به شناسایی این سرطان نیست. برای ارزیابی عملکرد این مدل بهتر است که از precision و recall استفاده کنیم.
- سناریو دوم: حال فرض کنید می خواهیم عملکرد مدلی را ارزیابی کنیم که با توجه به سابقه کاربر پیش بینی می کند که آیا کاربر بر روی یک تبلیغ به خصوص کلیک می کند یا خیر. برخلاف سناریو پیشین، TN در این سناریو مقدار زیادی است و تاثیر آن را فقط در accuracy می توانیم ببینم. بنابراین این معیار از دیگر معیار ها برای ارزیابی بهتر است.
- سناریو سوم: برای ارزیابی یک موتور جستجو معیار های accuracy و precision به اندازه recall برای این تسک مناسب نیستند. دلیل این ادعا این است که سایت های مرتبط با یک کوئری بسیار کمتر از سایت های غیر مرتبط است. و از انجایی که تمرکز recall تنها بر روی اسناد بازگردانده شده است، این معیار مناسب تر است.

سوال سوم:

سوال سوم:

$$info(D) = -\sum_{i=1}^{n} p_i \log_2(p_i)$$

$D_{\text{بیماری قلبی}} = [2,-2]$

$$info_{\text{بیماری قلبی}}(D) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1$$

$$info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times info(D_j)$$

$$info_{\text{درد سینه}}(D) = \frac{3}{4}\left(-\frac{2}{3}\log\frac{2}{3} - \frac{1}{3}\log\frac{1}{3}\right) + \frac{1}{4}\times\left(-1\log_2 1\right) = 0.69$$

0.91

$$gain(\text{درد سینه}) = 1 - 0.69 = 0.31$$

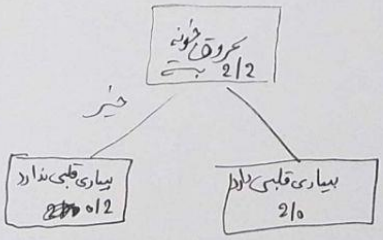
$$info_{\text{گردش خون}}(D) = \frac{2}{4}\times\left(0 - \frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2}\right) + \frac{2}{4}\times 1 = 1$$

$$gain(\text{گردش خون}) = 1 - 1 = 0$$

$$info_{\text{عروق خونی}}(D) = \frac{2}{4}\times 0 + \frac{2}{4}\times 0 = 0$$

$$gain(\text{عروق خونی}) = 1 - 0 = 1$$

با توجه به gain های بدست آمده عروق خونی بسته بیشترین وزن را داشته و به عنوان ریشه انتخاب می‌شود.



سوال چهارم:

سوال چهارم:

$info_{\text{دوست داشتن سریال کلاه قرمزی}}(D) = -\frac{3}{7}\log_2\frac{3}{7} - \frac{4}{7}\log\frac{4}{7} = 0.98$

$info_{\text{دوست داشتن پاپ کرن}}(D) = \frac{4}{7}\left(-\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4}\right) + \frac{3}{7}\left(-\frac{2}{3}\log\frac{2}{3} - \frac{1}{3}\log\frac{1}{3}\right)$
$= 0.85$

$gain(\text{دوست داشتن پاپ کرن}) = 0.14$

$info_{\text{دوست داشتن آب گازدار}}(D) = \frac{4}{7}\left(-\frac{3}{4}\log\frac{3}{4} - \frac{1}{4}\log\frac{1}{4}\right) + \frac{3}{7} \times 0 = 0.47$

$gain(\text{دوست داشتن آب گازدار}) = 0.98 - 0.47 = 0.51$

از آن جایی که سن یک متغیر عددی است، باید یک threshold برای آن انتخاب کنیم.

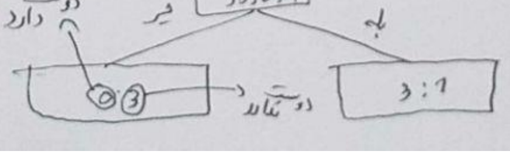
		15			44		
7	12	|	18	35	38	|	50	83
−	−		+	+	+		−	−

هر دو سن 44 و 15 یک شرایط خواهند داشت. با این حال 15 را به عنوان مرز انتخاب می کنیم.

$info_{\text{سن} < 15}(D) = \frac{2}{7} \times (-1\log 1) + \frac{5}{7}\left(-\frac{3}{5}\log\frac{3}{5} - \frac{2}{5}\log\frac{2}{5}\right) = 0.7$

$gain(\text{سن} > 15) = 0.29$

با توجه به gain های به دست آمده، درخت اولیه شکل زیر است:



ادامه ۴:

برای شاخه «بله» نیز نیاز به گره دیگری داریم

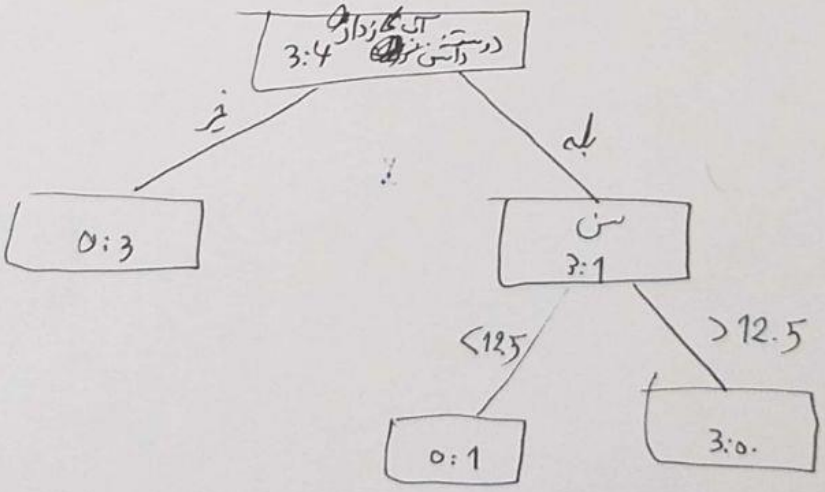
$$info_{\text{دوست داشتن کلاه قرمزی}} = -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4} = 0.81$$

$$info_{\text{دوست داشتن پاپ‌کرن}} = \frac{2}{4}\left(-\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2}\right) + \frac{2}{4}\times(-1\log 1) = 0.5$$

$$gain(\text{دوست داشتن پاپ‌کرن}) = 0.81 - 0.5 = 0.31$$

$$info_{\text{سن}>13} = \frac{1}{4}\times 0 + \frac{3}{4}\times 0 = 0$$

$$gain(\text{سن}>13) = 0.81$$



سوال پنجم:

سؤال پنجم: شاخص جینی معیاری برای اندازه‌گیری ناخالصی ویژگی است که درجه‌ای که یک متغیر به اشتباه دسته‌بندی شده باشد را به دست می‌آورد. مقدار صفر نشان دهنده این است که تمامی عناصر به یک کلاس تعلق دارند درحالی که مقدار 1 نشان دهنده این است که تمامی عناصر به طور رندوم بین کلاس های مختلف پخش شده اند.

$$Gini = 1 - \sum_{i=1}^{n} (P_i)^2$$

بیانگر نشان دهنده احتمال این که یک شئ خاص در کلاس خاصی طبقه‌بندی گردد.

برای مثال درخت سؤال چهارم را در نظر بگیرید.

$P(\text{آب گازدار} = +) = \frac{4}{7}$ $P(\text{آب گازدار} = -) = \frac{3}{7}$

$P(\text{سریال کلاه قرمزی} = + \mid \text{آب گازدار} = +) = \frac{3}{4}$

$P(\text{سریال کلاه قرمزی} = - \mid \text{آب گازدار} = +) = \frac{1}{4}$

$Gini = 1 - (\frac{9}{16} + \frac{1}{16}) = \frac{3}{8}$

$P(\text{سریال کلاه قرمزی} = + \mid \text{آب گازدار} = -) = 0$

$P(\text{سریال کلاه قرمزی} = - \mid \text{آب گازدار} = -) = 1$

$Gini(\text{آب گازدار} = -) = 1 - (0^2 + 1^2) = 0$

$Gini(\text{آب گازدار}) = P(\text{آب گازدار} = +) \times Gini(\text{آب گازدار} = +) + P(\text{آب گازدار} = -) \times Gini(\text{آب گازدار} = -) = \frac{4}{7} \times \frac{3}{8} + \frac{3}{7} \times 0 = \frac{3}{14}$

سوال ششم:

Overfitting زمانی رخ میدهد که مدل ما بر روی داده های آموزشی بیش از اندازه آموزش داده شود و یا مدلی که برای تسک طراحی شده است از پیچیدگی بالایی بر خوردار است. نتیجه آن دقت بالا در داده آموزش و دقت پایین در داده تست است. برای مثال انتخاب یک شبکه بسیار عمیق برای یک تسک ساده می تواند به این مشکل بخورد یا قرار ندادن limit بر روی عمق یک درخت تصمیم گیری نیز می تواند این مشکل را ایجاد کند.

برای جلوگیری از این مشکل می توان کار های زیر را انجام داد:

افزایش داده: از دلایل این مشکل داده کم است. افزایش تنوع در مجموعه داده آموزش می تواند این مشکل را تا حد خوبی برطرف کند.

ساده سازی مدل: راه دیگر ساده سازی ساختار مدل مورد استفاده است. برای مثال اگر این مدل یک شبکه عصبی است، می توان تعداد لایه های آن را کاهش داد یا اگر یک درخت تصمیم گیری است می توان عمق آن را کاهش داد.

Dropout: در این روش با در نظر گرفتن یک احتمال برای هر لایه از یک شبکه عصبی، برخی از لایه ها را در زمان آموزش نادیده می گیریم.

Regularization: تجربه ثابت کرده است که هر چه قدر اندازه وزن های یک مدل کوچکتر باشد، آن مدل ساده تر است. برای همین می توان از مکانیزمی استفاده کرد که در هنگام آموزش، مدل را برای داشتن وزن های بزرگ تنبیه کند. از این روش ها می توان به l1 و l2 regularization اشاره کرد.

$$Loss = Error(y, \hat{y}) + \lambda \sum_{i=1}^{N} |w_i|$$

Loss function with L1 regularisation

$$Loss = Error(y, \hat{y}) + \lambda \sum_{i=1}^{N} w_i^2$$

Loss function with L2 regularisation